RESEARCH ARTICLE OPEN ACCESS

Applications of Natural Language Processing

Ashu Jain[1] and Dhyanendra Jain[2]
1 ajainashu@gmail.com
2 dhyanendra.jain@gmail.com
Dr. Akhilesh Das Gupta Institute of Technology and Management,New Delhi, India

ABSTRACT: Natural language processing(NLP) is a branch of Artificial Intelligence (AI) through which computers can understand and analyze the human language. The communication between humans and computer is made possible by utilizing the power of Machine Learning. NLP offers various applications like text classification and summarization, speech recognition, character recognition, sentimental analysis, similarity detection, question answering and many more. In this paper, we have focused on three applications that are text summarization, optical character recognition(OCR) and similarity/duplicacy detection. Text summarization shortens the length of the document by concentrating on the main points only. OCR recognizes the characters present in the images. Similarity detection is used to detect the similarity between the two documents.

Keywords: Natural Language Processing, Text summarization, Optical Character Recognition, Duplication checker

Date of Submission: 18-05-2020 Date of Acceptance: 03-06-2020

I. INTRODUCTION

Natural language processing (NLP), one of the most invigorating fragments of AI is good to go to control the way where we talk with the outside world. NLP uses computational what's progressively, a logical strategy to separate the human language to energize relationship with PCs using conversational language. Subtopics right now natural language understanding of the wellsprings of information made by individuals, and natural language age, to focus on making natural language accounts.

The most standard approaches to manage NLP send Machine Learning. Natural language processing is improving human-PC conversations at the most extraordinary levels with applications or systems like Google Duplex, which can go about as an administrator to perform endeavors like creation haircut courses of action through phone by talking with individuals.

NLP is the voice behind Siri and Alexa, comparatively, customer help chatbots outfit the force of NLP to drive changed responses in web business, social protection and business utilities. A bit of the more pervasive usages of NLP today join remote aides, idea examination, customer help, and translation.

As development continues creating and advance, future use of NLP will be more customers masterminded. For example, remote partners will have the alternative to answer altogether increasingly befuddled requests reviewing the recommendations nearby the demanding hugeness of the inquiry asked. (Q: How is the atmosphere today? A: Rainy, you will require an umbrella).

Later on times to come, associations will have the alternative to offer a lot of master customer organizations, acknowledge summons right and uplift issues to certified people.

The utilization of NLP isn't restricted to appreciating customer questions or giving adjusted shopping/prosperity admonishment yet has formed into a continuously mechanical assistance of sorts. In the current events, NLP can be set up to give an overview of goofs, if one uses NLP to ask "what's new with my framework?" In the Future, NLP will be prepared to appreciate the customer's real objective like he needs his framework fixed for a passageway.The future with NLP is energizing as advances in NLP will permit humanity to move center from the inquiries to the outcomes. It will be a goliath jump when NLP can comprehend the client's information and gives progressively complex arrangements noting the client's genuine purpose.

This paper focuses on the real life solutions leveraging the power of NLP combined with machine learning.

The rest of the paper isarranged in the following manner: Section II describes the related work done in field of NLP. Section II gives the overview of the text summarization. Section IV discusses the Optical Character Recognition (OCR). Section Vexplains the Similarity duplication detection. In section VI, conclusion and directions of future work are provided.

II. RELATED WORK

There are various models proposed in the literature for the applications of NLP. Few important studies have been discussed here. Strobelt et al [1]overviewed and tried the adequacy of basic text featuring procedures, both separately and in blend, to find how to boost jump out impacts while limiting visual obstruction between strategies. They examined the upsides and downsides of various mixes as a structure rule to pick text featuring procedures for text watchers. Lin et al [2] proposed POME (Pattern-based Opinion MinEr), a methodology that uses characteristic language parsing and design coordinating to characterize Stack Overflow sentences alluding to APIs as indicated by seven perspectives (e.g., execution, ease of use), and to decide their extremity (positive versus negative). Their examination showed that POME displays a higher exactness than a best in class procedure (Opiner), as far as both conclusion angle distinguishing proof and extremity evaluation. Raphal et al [3] clarified the review of the different procedures in abstractive text summarization. It incorporates information handling, word installing, essential model engineering, preparing, and approval process.Kolal et al [4] presented a generative probabilistic OCR model that portrays a start to finish process in the loud channel system, advancing from age of genuine content through its change into the loud yield of an OCR framework. Lopresti [5] proposed for estimating the effect of recognition mistakes on the phases of a standard text investigation pipeline: sentence limit identification, tokenization, and grammatical feature labeling. Our approach figures mistake order as an advancement issue resolvable utilizing a various leveled dynamic programming approach. BilenkoandMooney [6] presented a system for improving duplicacy detection utilizing trainable proportions of textual similarity. They proposed to utilize learnable text separation capacities for each database field, and show that such measures are fit for adjusting to the particular thought of similarity that is fitting for the field's space. Basit and Jazrabek [7] answered for the issue of identifying some essential, yet helpful, sorts of configuration level similarities for example, gatherings of exceptionally comparable classes or records. They depicted the strategy for auxiliary clone detection, a model apparatus called Clone Miner that actualizes the strategy and trial results.

III. TEXT SUMMARIZATION

Text summarization alludes to the strategy of shortening long bits of text. Automatic text summarization is a typical issue in AI and NLP.

Moved by the cutting edge mechanical developments, information is to this century what oil was to the past one. Today, our reality is parachuted by the get-together and spread of colossal measures of information.

Actually, the International Data Corporation (IDC) ventures that the aggregate sum of advanced information flowing every year around the globe would grow from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. That is a great deal of information!

With such a major measure of information coursing in the advanced space, there is have to create AI calculations that can naturally abbreviate longer texts and convey precise outlines that can smoothly pass the proposed messages.

Moreover, applying text summarization decreases understanding time, quickens the procedure of inquiring about for data, and builds the measure of data that can fit in a region.

Algorithm for text summarization

As a rule, text summarization in NLP is treated as an administered AI issue (where future results are anticipated dependent on given information).

Ordinarily, here is the manner by which utilizing the extraction-based way to deal with outline texts can work:

1. Acquaint a technique with remove the justified keyphrases from the source archive. For instance, you can utilize grammatical form labeling, words groupings, or other phonetic examples to recognize the keyphrases.
2. Accumulate text records with decidedly marked keyphrases. The keyphrases ought to be good to the stipulated extraction method. To build exactness, you can additionally make adversely marked keyphrases.
3. Train a double AI classifier to make the text summarization. Some of the highlights you can utilize include:
    - Length of the keyphrase
    - Frequency of the keyphrase
    - The most repeating word in the keyphrase
    - Number of characters in the keyphrase
4. At last, in the test expression, make all the keyphrase words and sentences and convey out arrangement for them.

IV. OPTICALCHARACTER RECOGNITION (OCR)

OCR, or optical character recognition, is one of the most punctual tended to PC vision errands, since in certain viewpoints it doesn't require profound learning. In this way, there were distinctive

OCRusage even before the profound learning blast in 2012, and some even gone back to 1914.

This makes numerous individuals think the OCR challenge is "understood", it is never again testing. Another conviction which originates from comparable sources is that OCR doesn't require profound learning, or as it were, utilizing profound learning for OCR is a needless excess.

Any individual who rehearses PC vision, or AI when all is said is done, realizes that there is nothing of the sort as a comprehended assignment, and this case isn't unique. Despite what might be expected, OCR yields awesome outcomes just on quite certain utilization cases, yet all in all, it is still considered as trying.

Also, it is valid that there are acceptable answers for certain OCR errands that don't require profound learning. Be that as it may, to truly step forward towards better, increasingly broad arrangements, profound learning will be required.

A few characteristics of the OCR undertakings:

- Text thickness: on a printed/composed page, content is thick. Be that as it may, given an picture of a road with a solitary road sign, content is meager.
- Structure of content: message on a page is organized, for the most part in exacting lines, while content in the wild might be sprinkled all over the place, in various revolutions.
- Fonts: printed text styles are simpler, since they are increasingly organized then the boisterous written by hand characters.
- Character type: content may come in various languages which might be very not quite the same as one another. Furthermore, structure of content might be unique in relation to numbers, for example, house numbers and so on.
- Artifacts: unmistakably, open air pictures are a lot noisier than the agreeable scanner.
- Location: a few undertakings incorporate trimmed/focused content, while in others, content may be situated in irregular areas in the picture.

V. SIMILARITY DUPLICATION DETECTION

In this part of the paper, we are concentrating on a text similarity checker that can be utilized from numerous points of view. There is such a great amount of substance on the web and ordinarily the content is like the numerous others. Utilizing the similarity identification device the substance on different destinations can be overseen. Duplication and Plagiarism can likewise be distinguished and can be valuable to keep up the honesty of online records.

VI. CONCLUSION AND FUTURE SCOPE

In this paper we have concentrated on three applications that are text summarization, optical character recognition(OCR) and similarity/duplicacy detection. Text outline abbreviates the length of the report by focusing on the primary concerns as it were. OCR perceives the characters present in the pictures. Similarity detection is utilized to recognize the similarity between the two reports.

In this paper, we have discussed only three applications of natural language processing that are text summarization, optical character recognition(OCR) and similarity/duplicacy detection. In future studies, other applications of NLP can be explored which sentimental analysis, speech recognition, chatbots, spelling checkers and many more.

References

[1]. H. Strobelt, D. Oelke, B. C. Kwon, T. Schreck and H. Pfister, "Guidelines for Effective Usage of Text Highlighting Techniques," in *IEEE Transactions on Visualization and Computer Graphics, vol. 22,* no. 1, pp. 489-498, 31 Jan. 2016

[2]. B. Lin, F. Zampetti, G. Bavota, M. Di Penta and M. Lanza, "Pattern-Based Mining of Opinions in Q&A Websites," *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE),* Montreal, QC, Canada, 2019, pp. 548-559.

[3]. N. Raphal, H. Duwarah and P. Daniel, "Survey on Abstractive Text Summarization," *2018 International Conference on Communication and Signal Processing (ICCSP),* Chennai, 2018, pp. 0513-0517.

[4]. O. Kolak, W. Byrne and P. Resnik, "A Generative Probabilistic OCR Model for NLP Applications", *2 Proceedings of HLT-NAACL 2003 Main Papers* , pp. 5 . Edmonton, May-June 2003

[5]. Lopresti, D. Optical character recognition errors and their effects on natural language processing. *IJDAR 12, 141–151 (2009).*

[6]. M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining,* August 2003, pages39–48

[7]. H. A Basit and S. Jarzabek, "Detecting higher-level similarity patterns in programs", *ACM SIGSOFT Software Engineering Notes,* September 2005.